

Delay asymptotics in the $GI/GI/1$ PS queue

Bert Zwart

Department of Mathematics and Computer Science
Eindhoven University of Technology

Center for Mathematics and Computer Science
Advanced communication networks

Joint work with: Sem Borst (Bell-labs,CWI), Michel Mandjes (CWI),
Sindo Núñez (CWI,TUE), Dennis van Ooteghem (TUE).

The basic model: $GI/GI/1$ PS

- Jobs arrive according to a renewal process;
i.e. interarrival times $\{A_i, i \in \mathbb{N}\}$ are i.i.d.;
- Job sizes $\{B_i, i \in \mathbb{N}\}$ are i.i.d.;
- When there are n jobs in the system, each of them is served at rate $1/n$.
- Stability criterion: $\rho := \mathbb{E}\{B\}/\mathbb{E}\{A\} < 1$.

Motivation

- PS queues were originally intended to analyze time-sharing in computer networks (see e.g. Kleinrock).
- Nowadays: PS is commonly used to model elastic traffic in communication networks.

Elastic traffic: Traffic that is not subject to tight delay requirements (e.g. data opposed to voice).

- Example: A link with long-lived identical TCP flows (elephants). By the nature of TCP, every data flow eventually gets the same share of the bandwidth, i.e. the server divides its attention equally over all flows.
- *Warning:* PS is an idealization of TCP (the above fair-sharing property is assumed to be established instantaneously)

Main performance measures

Q := Steady-state queue length

V := Sojourn time of an arbitrary customer

$V(\tau)$:= Sojourn time of an arbitrary
customer with job size τ .

Well-known results for M/G/1 PS:

- Q has a geometric distribution:

$$\mathbb{P}\{Q = n\} = (1 - \rho)\rho^n.$$

- The mean of $V(\tau)$ is linear in τ :

$$\mathbb{E}\{V(\tau)\} = \frac{\tau}{1 - \rho}.$$

- Fluid and diffusion limits for $GI/GI/1$: Gro-moll, Puhá, Stolyar, Williams.

Main topic of this talk: Long sojourn times

- $\mathbb{P}\{V > x\}, x \rightarrow \infty.$
- Large deviations techniques.
“Unlikely events happen in the most likely way”
- Long sojourn times can be caused by
 1. The job of the tagged customer;
 2. The jobs present at the arrival of the tagged customer;
 3. The jobs arriving during the sojourn time of the tagged customer.

Overview

- Heavy tails
- Light (exponential) tails
- A mixed (two-class) scenario

Heavy tails

Theorem (for $M/G/1$ PS)

If $\mathbb{P}\{B > x\} \sim L(x)x^{-\nu}, \nu > 1$, then

$$\mathbb{P}\{V > x\} \sim \mathbb{P}\{B > x(1 - \rho)\}. \quad (1)$$

- First shown in Zwart & Boxma (2000). Extensions by Nunez-Queija (2002) (to \mathcal{IR}) and Jelenkovic & Momcilovic (2003) (to \mathcal{SC}).
- Extension to state dependent PS queues (e.g. $M/G/s$ PS) in Guillemin, Robert & Zwart (2004).
- Intuition: In the long run, a large customer is served at a reduced rate $1 - \rho$.

Therefore, (1) is sometimes called a **reduced service rate** (RSR) approximation.

- PS behaves superior to FIFO (which gives $O(x\mathbb{P}\{B > x\})$ behavior).

Cycle formula approach (BOZ04)

In all cases considered so far, proofs exploited information about the queue length (exact results or upper bounds).

Problem in e.g. $GI/GI/1$ PS and DPS: No results about queue length are known.

Different idea for the $GI/GI/1$ queue: Use the formula

$$\mathbb{P}\{V > x\} = \frac{1}{\mathbb{E}\{N\}} \mathbb{E}\left\{ \sum_{i=1}^N I(V_i > x) \right\}.$$

N is the number of customers in a busy cycle.

Heuristics

Recall that P is the length of a busy period.

Z01: $\mathbb{P}\{P > x\} \sim \mathbb{E}\{N\}\mathbb{P}\{B > x(1 - \rho)\}$.

Claim: A large ($O(x)$) busy period contains exactly one large customer indexed by i^* .

$$\begin{aligned}\mathbb{P}\{V > x\} &= \frac{1}{\mathbb{E}\{N\}}\mathbb{E}\left\{\sum_{i=1}^N I(V_i > x)\right\} \\ &= \frac{1}{\mathbb{E}\{N\}}\mathbb{E}\left\{\sum_{i=1}^N I(V_i > x)I(P > x)\right\} \\ &= \frac{1}{\mathbb{E}\{N\}}\mathbb{P}\{V_{i^*} > x; P > x\} \\ &\quad + \frac{1}{\mathbb{E}\{N\}}\mathbb{E}\left\{\sum_{i \neq i^*} I(V_i > x)I(P > x)\right\} \\ &\approx \frac{1}{\mathbb{E}\{N\}}\mathbb{P}\{P > x\} + o(\mathbb{P}\{B > x\}) \\ &\sim \mathbb{P}\{B > x(1 - \rho)\}.\end{aligned}$$

Main problem: Get precise information on the number of large customers conditionally upon $P > x$.

Multiclass extensions, e.g. DPS

DPS discriminates between various customer classes. There are weights g_i such that customers of class i are served with rate $g_i / \sum_j n_j g_j$, with n_j the number of customers in the system which belong to class j .

Theorem (BOZ04)

If $\mathbb{P}\{B_1 > x\} = L(x)x^{-\nu}$ and $\mathbb{P}\{B_j > x\} = o(\mathbb{P}\{B_1 > x\})$, $j > 1$, then

$$\mathbb{P}\{V_1 > x\} \sim \mathbb{P}\{B_1 > (1 - \rho)x\}$$

Recall that long sojourn times can be caused by

1. The job of the tagged customer;
2. The jobs present at the arrival of the tagged customer;
3. The jobs arriving during the sojourn time of the tagged customer.

Conclusion:

Effect 1 dominates in the heavy-tailed case.

Light tails

(In)famous!

Flatto (1997) for M/M/1

(arrival rate λ , mean service time μ^{-1}):

$$\mathbb{P}\{V > x\} \sim c x^{-5/6} e^{-\alpha x^{1/3}} e^{-\gamma x}, \quad x \rightarrow \infty,$$

where $\gamma := (\sqrt{\mu} - \sqrt{\lambda})^2$.

Interestingly, for the busy period P :

$$\mathbb{P}\{P > x\} \sim c' x^{-3/2} e^{-\gamma x}, \quad x \rightarrow \infty,$$

see Cox & Smith (1961), Palmowski & Rolski (2004).

Conjecture: V and P have the *same* logarithmic asymptotics...

Upper bound

Sojourn time is always shorter than residual busy period!

$$\{V > x\} \subseteq \{B_0 + W + A(x) > x\},$$

where

B_0 := The job of the tagged customer;

W := The amount of unfinished work of the jobs present at time 0;

$A(x)$:= The work generated by jobs arriving in $(0, x]$.

Recognize effect 1, 2, 3!

The usual Chernoff bound gives, for any $\nu > 0$,

$$\begin{aligned} \mathbb{P}\{V > x\} &\leq \mathbb{P}\{B_0 + W + A(x) - x > 0\} \\ &\leq \mathbb{E}\{e^{\nu B}\} \mathbb{E}\{e^{\nu W}\} \mathbb{E}\{e^{-\nu(x - A(x))}\}. \end{aligned}$$

Upper bound, ctd.

Define

$$\Phi_B(\nu) := \mathbb{E}\{e^{\nu B}\}, \Phi_A(\omega) := \mathbb{E}\{e^{\omega A}\}.$$

The asymptotic cgf of $A(x)$ is given by

$$\Psi(s) := \lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{E} e^{sA(x)} = -\Phi_A^{\leftarrow} \left(\frac{1}{\Phi_B(s)} \right),$$

Hence,

$$\limsup_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{E}\{e^{-\nu(x-A(x))}\} \leq \inf_{\nu > 0} \left(-\Phi_A^{\leftarrow} \left(\frac{1}{\Phi_B(\nu)} \right) - \nu \right).$$

Upper bound, ctd.

We call

$$\gamma := - \inf_{\nu > 0} \left(-\Phi_A^{\leftarrow} \left(\frac{1}{\Phi_B(\nu)} \right) - \nu \right);$$

optimizer: ν^* .

It is easily verified that $\mathbb{E}\{e^{\nu^*B}\} < \infty$ and $\mathbb{E}\{e^{\nu^*W}\} < \infty$. Hence

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{P}\{V > x\} \leq -\gamma,$$

as desired.

For M/M/1, we indeed have

$$\gamma = (\sqrt{\mu} - \sqrt{\lambda})^2.$$

Intuition: only effect 3 plays a role in logarithmic asymptotics...

Lower bound

Sketch of proof:

- Perform the following change of measure:
 - Leave the tagged job unchanged;
 - Leave the jobs present at time 0 unchanged;
 - Perform an *exponential* change of measure

$$\tilde{F}_A(dx) = F_A(dx) \frac{e^{\omega x}}{\Phi_A(\omega)};$$

$$\tilde{F}_B(dx) = F_B(dx) \frac{e^{\nu x}}{\Phi_B(\nu)}$$

in the interval $(0, x]$.

- Choose $\omega = \omega_\epsilon$, $\nu = \nu_\epsilon$ such that $\Phi_A(\omega)\Phi_B(\nu) = 1$, and the load under the new measure is $1 + \frac{\epsilon}{2}$.

Lower bound, ctd.

- Standard: $\mathbb{P}\{V > x\} = \mathbb{E}_\epsilon\{L_x I(V > x)\}$;
expectation under new probability measure.
Evidently, for any event S ,

$$\mathbb{P}\{V > x\} \geq \mathbb{E}_\epsilon\{L_x I(V > x) I(S)\}.$$

- Choose an appropriate S such that
 - On $\{V > x\} \cap S$, L_x can be bounded appropriately:

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{E}_\epsilon\{L_x \mid V > x, S\} \geq -\nu_\epsilon - \omega_\epsilon(1 + \epsilon);$$

- $\mathbb{P}_\epsilon\{V > x, S\}$ decays subexponentially.
- Let $\epsilon \downarrow 0$, then

$$-\nu_\epsilon - \omega_\epsilon(1 + \epsilon) \rightarrow -\gamma.$$

Lower bound, ctd.

In our proof:

$\mathbb{P}_\epsilon\{V > x, S\}$ decays subexponentially...

... because $\mathbb{P}_\epsilon\{V > x\}$ does.

Proof relies on Puha, Stolyar, Williams (2004): FLLN for overloaded PS queues.

$$\{Q(ut)/t\}_{u \geq 0} \rightarrow \{qu\}_{q \geq 0}$$

Hence:

$$\begin{aligned} \mathbb{P}_\epsilon\{V > x\} &= \mathbb{P}_\epsilon\left\{B_0 > \int_0^x \frac{1}{1 + \bar{Q}(u)} du\right\} \\ &\approx \mathbb{P}_\epsilon\left\{B_0 > \int_0^x \frac{1}{1 + q_\epsilon u} du\right\} \\ &\approx \mathbb{P}\{B_0 > \text{const} + (1/q_\epsilon) \log x\} \end{aligned}$$

Lower bound, ctd.

Additional condition is required:

For each constant $c > 0$, we have

$$\lim_{x \rightarrow \infty} \frac{1}{x} \log \mathbb{P}\{B_0 > c \log x\} = 0.$$

Excludes:

- extremely light tails, and
- distributions of bounded support.

Indeed, M/D/1 PS gives a different decay rate...

Conclusions

- For light-tailed service times, large sojourn times primarily occur due to effect 3: After time 0 work is arriving at rate $1(> \rho)$.
- For $M/G/1$ we formulate an asymptotically optimal importance sampling algorithm.
- Numerical experiments show that the algorithm is superior w.r.t. straightforward Monte-Carlo simulation.
- Flatto's exact asymptotic for $M/M/1$ behaves poorly as an approximation if ρ is large.
- Large deviations and heavy traffic limits cannot be interchanged!

A mixed scenario (BNZ03)

Consider the following $M/G/1$ PS queue:

- Customers of type 1 arrive according to a Poisson process with rate λ_1 and have an exponential (μ) service time distribution.
- Customers of type 2 arrive according to a Poisson process with rate λ_2 with $\mathbb{P}\{B_2 > x\} = L(x)x^{-\nu}$.
- Both customer types are served at rate $1/n$.

Main question:

How is the tail of $\mathbb{P}\{V_1 > x\}$ affected by class-2 customers?

- Finite waiting room K
- Infinite waiting room

Finite waiting room K

Theorem

$$\mathbb{P}\{V_1 > x\} \sim \frac{(1 - \rho)\rho_2^{K-1}}{1 - \rho^{K+1}} \mathbb{P}\{B_2^{res} > x/K\}^{K-1} e^{-\mu x/K}.$$

Intuition:

- $\{V_1 > x\}$ is caused by the presence of $K - 1$ heavy-tailed customers who remain in the system during the entire sojourn time of the tagged customer.
- Effect 2 dominates in this situation.
- What if $K = \infty$??

If $K = \infty$, we obtain subexponential behavior:

$$\mathbb{P}\{V_1 > x\} \geq e^{-c\sqrt{x \log x}(1+o(1))}.$$

PS is not as bad as FIFO but (conjecture) SRPT probably significantly better!

Main conclusions

What causes a long sojourn time?

- If a customer has a heavy-tailed service time distribution, effect 1 dominates.
- If all customers have light-tailed service time distributions, effect 3 dominates.
- If the tagged customer is light-tailed, but other customers are heavy-tailed, effect 2 dominates.